DOCUMENT RESUME

ED 065 593 TM 001 862

AUTHOR Randall, Robert S.
TITLE Contrasting Norm Referenced and Criterion Referenced Measures.
PUB DATE Apr 72
NOTE 11p.; Paper prepared for symposium of the Annual Meeting of the American Educational Research Association (Chicago, Illinois, April 1972)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Comparative Analysis; *Criterion Referenced Tests; *Norm Referenced Tests; *Test Construction; *Test Reliability; *Test Validity

ABSTRACT

Differences in design between norm referenced measures (NRM) and criterion referenced measures (CRM) are reviewed, and some of the procedures proposed on designing and evaluating CRM are examined. Differences in design of NRM and CRM are said to arise from the different purposes that underlie each measure. In addition, there are differences among criterion referenced tests, three cases of which are: (1) where items are sampled from a known universe, (2) where one item constitutes the set in question, and (3) where items are examples of a class of problems or tasks which cannot be well defined. Validation problems in CRM are discussed, and the need for developing new techniques, especially for case (3) CRTs, is pointed out. (DB)

ED 065593

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

CONTRASTING NORM REFERENCED AND CRITERION REFERENCED MEASURES

by

Robert S. Randall, Ph.D.
Southwest Educational Development Laboratory
Austin, Texas
March, 1972

TM 001 862

Prepared for American Educational Research Association Symposium entitled A Model for Estimating the Reliability and Validity of Criterion Referenced Measures, Chicago, Illinois, April, 1972

CONTRASTING NORM REFERENCED AND CRITERION REFERENCED MEASURES

Robert S. Randall

The use of traditional methods for designing and evaluating Criterion Referenced Measures (CRM) have increasingly come under attack in recent years. Cox and Vargus (1966) have discussed problems as early as 1966 that are encountered when classical methods developed for norm referenced measures are used to evaluate criterion referenced measures. Popham (1969) and Ivens (1970) have also published warnings and offered suggestions for different approaches. Kriewall (1969) examined these discussions at length and composed a model for curriculum design and management including criterion referenced measures and their evaluation. We will return later to Kriewall's work. Livingston (1972) has also examined these difficulties and has recently proposed a solution of his own. Before examining some of the procedures that have been proposed on designing and evaluating criterion referenced measures, let us review briefly some of the differences in design between norm referenced measures and criterion referenced measures that cause difficulty in validating the latter.

Differences in Design of NRM vs. CRM

Most of the differences in design have to do with different purposes that underly the measure. NRM design assumes a trait or ability is present in varying degrees in different individuals. The attempt is

to design a measure that will separate these individuals in terms of scores on the test which measure that trait or ability. Thus, the test items must constitute a homogeneous set, all of which measure some degree of the ability in question. While a CRT may be a homogeneous set of items, the concern is to measure some defined level of development or mastery of some specified class of problems or tasks. Whether subjects are able or unable to perform well on the test items is of little concern to the designers, although it is hoped that, after instruction, a given set of subjects with given prerequisite development will be able to do well on the test. Thus, a CRT may or may not contain a homogeneous set of items. In fact, as will be demonstrated later, one item may for all practical purposes constitute an entire CRT. Hence, one set of items that appears to be one test may be treated as several one-item tests. This has implications for the attention given in norm referenced tests construction under the topic of reliability which is related to internal consistency of measurement on a given test. We'll return to this problem later.

Another difference in constructing NRM vs. CRM is the difficulty index of items. In a NRT this difficulty level of each item is of great concern and must correspond to and aim at a given population norm. Typically, items are constructed that have something close to a .5 difficulty index so that about half the population at which the test is normed or aimed will get an item correct. Items above .75 or below .25 in difficulty index are usually discarded because they are too easy or too difficult for the population in question. In constructing a CRT,

difficulty is not a function of a population, but rather a function of development or mastery level which is specified by the curriculum objectives. Therefore, items which everyone of a given population passed or failed might be included in the test since the object is to measure mastery or proficiency in some area at a defined level.

Another difference resulting from the different purposes of the tests has to do with the discrimination power of items in the test. Since the assumption of the NRM is that differences exist among individuals in ability or acquisition of a trait, the test must be designed to demonstrate that items do in fact discriminate between those who have the ability in greater degrees and those who have it in lesser degrees. Thus, variance on the test is exceedingly important since differences are assumed to exist among the subjects. Evidence is gathered to indicate that subjects who do well on the tests as a whole, do well on the more difficult items or at least better than those who do poorly on the test. Every item is expected to have this kind of discrimination power to some extent. That is, those who tend to do well on the total test should tend to do better on each item than those who did poorly on the whole test. Items that fail in this respect are discarded. In contrast, a CRT item may or may not discriminate. If it does, it is fine, but if it should not, it is not cause on that basis alone to discard the item, as is true in the case of the NRT. Again, the fact that all subjects may score very close to perfect on the tests will cause the variance to be extremely low and may result in a low discrimination index for an item that is entirely an artifact

of the low variance. Of course, if a CRT has a homogeneous set of items measuring a class of problems, discrimination power of each item may be of concern, but the way to determine it is the question. It is clear that use of classical statistical methods designed to estimate discrimination power of norm referenced items are of little value in establishing discrimination power of criterion referenced items. Thus, the manner in which designers construct items for criterion referenced tests is greatly different from that in which norm referenced items are designed, in that fine degrees of difference and discrimination power are not of primary concern to CRT designers.

Differences among CRTs

There are, of course, differences among criterion referenced tests. There appear to me to be at least three cases of CRTs. Case 1 is where items are sampled from a known universe. Kriewall (1969) has described at length this method which is based on specifying a well defined set of problems or tasks that constitutes a known, finite universe of test items from which random samples are drawn without replacement, thus creating a finite number of tests with some given number of items each. An example of such a well defined set, which Kriewall calls "specified content objectives" (SCO), is addition of any two integers or single digit numbers from 0 through 9 inclusively. Another example would be recognition of all three letter words beginning with the letter N.

A case 2 is where one item constitutes the set in question. In other words, the class of tasks is a one-element set. Examples of such an item include riding a bicycle ten feet without falling or

touching the ground (some might argue that proficiency could be measured better with a criterion of two out of three attempts being successful) and playing a piano solo to some criterion of proficiency.[1]

Case 3 items are those which are examples of a class of problems or tasks which cannot be well defined, although they can be described or defined rather accurately. The set may be well defined in the sense that a given item can be determined to be in or not in the set, but the number of items possible is not known. The difference between case 3 and case 1 is that the finite universe of items is not known to the test items writers and thus the test items cannot constitute a random sample. Rather they are an illustrative set of items that are examples of the class in question. In fact, only one item may be used because of practical considerations, but it is assumed that others of the same class could be constructed. Examples of such items are recognizing the meter of a poem, recognizing the concept of dependence, or discriminating size, color or shape. SEDL's experience (and many others') has been with case 3 items almost exclusively.

Another difference among criterion referenced tests (as is true of NRTs) is the response mode that is used. Kriewall (1969) suggests that a constructed response is preferable since guessing errors are

---

[1] I have a feeling that Kriewall might argue that the inclusion of case 2 items in an instructional system would work for an inefficiency toward too many test items that eat away time of instruction. While this practical argument may be valid, there do appear to exist many examples of case 2 test items.

eliminated. While this is true in mathematics problems, on which his model was developed, in other content areas the reliability of scoring becomes an overpowering matter of concern, possibly more important than chance guessing errors. For example, writing a theme is a very sophisticated, constructed response with well known difficulties in reliability of scoring. Thus, some CRTs use alternate choice response modes which may be dichotomous or have more than two choices available. The choice of responses mode, however, affects the confidence one has in the validation procedure that is used.

Validation Problems in CRM

One concept of concern to test constructers is reliability. Reliability is discussed in text books on norm referenced measures in terms of internal consistency and stability or test-retest reliability. Internal consistency estimates of the reliability of a test usually look at relationships between the variance of responses to each item and the variance of total test response scores. As previously noted, such a concept is not of primary concern to CRT designers, but ever if it were, the usual methods are totally inadequate, since the number of items is usually small and alpha indexes are a function of numbers of items, to a great extent. If an internal consistency measure is high on a CRT, one may be pleased, but if it is low, one need not be displeased.

Stability of criterion referenced items from one measuring time to another are exceedingly important. However, as the following data illustrate, the typical methods are not valid because they are too much dependent on large numbers of items.

Consider the following results of repeating an alternate response mode test of 5 items with one subject:

	Test A	Test A[1]
	Items	Items
	1 Right	1 Wrong
	2 R	2 W
	3 W	3 R
	4 W	4 R
	5 R	5 R
Total Score	3	3

If similar results occurred with 50 other subjects the test-retest r would be perfect (1.0). If this were a NRT the results shown could not occur over a large number of subjects, since item analysis on difficulty index and discrimination power would likely have eliminated such items. Additionally, since a large number of items would reveal such erroneous results of guessing more readily, and one could guard against such a trap in test-retest reliability estimation. This is why larger numbers of items are used on NRTs and confidence is low on tests with small numbers of items. But, the nature of CRTs and their use demands small numbers of items and sometimes examination of stability of each item. Therefore, a different method is needed to estimate this reliability.

Validity of Criterion Referenced Tests has most often been established by some form of content validity. Kriewall, in fact, assumes validity because of the nature of his specified content objects in that test items are a random sample of the universe of such problems. Hence, he gives no discussion to validity. It seems apparent that

validity is not such a problem for case 1 and case 2 tests as it is for case 3. Similar problems to those discussed in estimating reliability of criterion referenced measures with procedures that were developed for norm referenced measures apply as well to validation procedures for criterion referenced tests. Construct validity, established by considering the range of total scores on a test from a number of subjects and compared in a correlation matrix with scores the same subjects made on other tests that are presumed to measure the same construct, depends heavily on comparisons of rank order of total scores being relatively stable between the tests. Since the number of items of a criterion referenced measure is usually smaller and the variance may be very small, such comparison between criterion referenced tests and the other test scores are not very promising, because the resulting score may be an artifact of the low variance on the CRT. The same is true of methods used in predictive validity. However, the concepts of construct and predictive validity may be very important to CRT designers, especially those who worked with case 3 items.

We have attempted to review the situation that faces those who wish to validate CRTs and show the need for developing new techniques, especially for case 3 type CRTs on which many curriculum designers are relying. The need exists for new techniques to be developed that will estimate reliability where test-retest stability is of concern and also to provide estimates for construct and predictive validity. Oakland (1972) examines some of the techniques in more detail that have been proposed and used. Following Oakland's paper, Edmondston (1972)

demonstrates how some techniques were proposed and evaluated in arriving at the model which will be subsequently presented (Edmondston, Randall, and Oakland (1972).

BIBLIOGRAPHY

Cox, R. C., & Vargas, Julie S. "A Comparison of Item Selection Techniques for Norm-Referenced and Criterion-Referenced Tests." Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, February, 1966.

Edmonston, Leon P. "A Review of Attempts to Arrive at More Suitable Evaluation Models: An Introspective Look." A paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, February, 1966.

Ivens, Stephen H. "A Pragmatic Approach to Criterion-Referenced Measures." Paper presented at a joint session of the annual meetings of the American Educational Research Association and the National Council on Measurement in Education, Chicago, Illinois, April, 1972.

Kriewall, T. E. "Application of Information Theory and Acceptance Sampling Principles to the Management of Mathematics Instruction." Technical Report No. 103, Wisconsin Research and Development Center, Madison, October, 1969.

Livingston, Samuel A. "A Classical Test-Theory Approach to Criterion-Referenced Tests." Paper presented at the American Educational Research Association, Chicago, Illinois, April, 1972.

Livingston, Samuel A. "The Reliability of Criterion-Referenced Measures." Report No. 73. The Center for the Study of Social Organization of Schools, The Johns Hopkins University, July, 1970.

Oakland, Thomas D. "An Evaluation of Available Models for Estimating the Reliability and Validity of Criterion Referenced Measures." A paper presented at the American Educational Research Association, Chicago, Illinois, April, 1972.

Popham, James W. & Husek, T. R. "Implications of Criterion-Referenced Measurement." Journal of Educational Measurement, Vol. 6, No. 1, Spring, 1969.

Randall, Robert S., Edmonston, Leon P., & Oakland, Thomas D. "A Model for Estimating the Reliability and Validity of Criterion Referenced Measures." Paper presented at the American Educational Research Association, Chicago, Illinois, April, 1972.